Computing Robust Leverage Diagnostics when the Design Matrix Contains Coded Categorical Variables

Kjell Konis

January 23, 2013

Abstract

For a robust leverage diagnostic in linear regression, Rousseeuw and van Zomeren [1990] proposed using *robust distance* (Mahalanobis distance computed using robust estimates of location and covariance). However, a design matrix X that contains coded categorical predictor variables is often sufficiently sparse that robust estimates of location and covariance cannot be computed. Specifically, matrices formed by taking subsets of the rows of X are likely to be singular, causing algorithms that rely on subsampling to fail. Following the spirit of Maronna and Yohai [2000], we observe that extreme leverage points are extreme in the continuous predictor variables. We therefore propose a robust leverage diagnostic that combines a robust analysis of the continuous predictor variables and the classical definition of leverage.

1 Background

We consider linear regression models of the form

$$y_i = x_{i1}^\top \beta_1 + x_{i2}^\top \beta_3 + x_{i2}^\top \beta_3 + e_i \qquad (i = 1, \ldots, n) \tag{1}$$

where $x_{i1} \in \mathbb{R}^{p_1}$ contains coded categorical predictor variables, $x_{i2} \in \mathbb{R}^{p_2}$ contains continuous predictor variables and the elements of $x_{i3} \in \mathbb{R}^{p_3}$ are each products of at least one element of x_{i1} and at least one element of x_{i2}. Let X_k be the matrix with i row $x_{ik}^\top$ for $k = 1, 2, 3$ so that the design matrix $X = [X_1\ X_2\ X_3]$. The dimension of X is $n \times p$ where $p = p_1 + p_2 + p_3$.

Two classical leverage measures are the diagonal elements of the hat matrix (the *hat values*)

$$h_i = H_{ii} = x_i^\top (X^\top X)^{-1} x_i \qquad (i = 1, \ldots, n) \tag{2}$$

where $x_i^\top = (x_{i1}^\top\ x_{i2}^\top\ x_{i3}^\top)$ is the i row of X and the *Mahalanobis distance* (MD)

$$\mathrm{MD}_i = \sqrt{(x_i^* - T(X^*))^\top C(X^*)^{-1}(x_i^* - T(X^*))} \tag{3}$$

where $T(X^*)$ is the arithmetic mean, $C(X^*)$ is the sample covariance matrix and X^* is identical to X except that the constant column has been removed (if present in X). When X does contain a constant column, these two measures are related by

$$h_i = \frac{(MD_i)^2}{n-1} + \frac{1}{n}. \tag{4}$$

2 Robustification

Let $\{T^{(rob)}, C^{(rob)}\}$ be a robust estimator of location and covariance where the final estimate is a weighted mean and a weighted covariance matrix with weights $w = (w_1, \ldots, w_n)^\top$, $w_i \in \{0, 1\}$. The covariance estimator $C^{(rob)}$ can additionally be rescaled by a factor c. The Fast MCD of Rousseeuw and van Driessen [1999] is one such estimator. The final robust estimate of location is

$$T^{(rob)}(X_2) = \frac{X_2^\top w}{\sum_{i=1}^n w_i}$$

and the final robust estimate of covariance is

$$C^{(rob)}(X_2) = \frac{c}{(\sum_{i=1}^n w_i) - 1}(X_2 - M)^\top \operatorname{diag}(w)\,(X_2 - M)$$

where M is an $n \times p_2$ matrix with rows $[T^{(rob)}(X_2)]^\top$.

We then observe that the following modification of X_2

$$\tilde{X}_2 = \sqrt{\frac{c(n-1)}{(\sum_{i=1}^n w_i) - 1}}\, W(X_2 - M) + M. \tag{5}$$

yields

$$T(\tilde{X}_2) = T^{(rob)}(X_2) \quad \text{and} \quad C(\tilde{X}_2) = C^{(rob)}(X_2). \tag{6}$$

Our idea is to form the *modified design matrix* $\tilde{X} = [X_1\ \tilde{X}_2\ \tilde{X}_3]$ where $\tilde{X}_3$ is formed as X_3 but using the values in $\tilde{X}_2$ in place of those in X_2. We then define the *robust hat value* to be

$$h_i^{(rob)} = x_i^\top (\tilde{X}^\top \tilde{X})^{-1} x_i \qquad (i = 1, \ldots, n) \tag{7}$$

and the *robust distance* to be

$$\mathrm{RD}_i = \sqrt{(x_i^* - T(\tilde{X}^*))^\top C(\tilde{X}^*)^{-1}(x_i^* - T(\tilde{X}^*))}. \tag{8}$$

3 Discussion

When the linear regression model contains only an intercept term and continuous predictor variables, $X^* = X_2$, $T(\tilde{X}^*) = T^{(rob)}(X_2)$ and $C(\tilde{X}^*) = C^{(rob)}(X_2)$ so that the quantity defined in equation 8 is equivalent to the robust distance given in Rousseeuw and van Zomeren [1990]. Hence, we call this quantity *robust distance* as well.

When $p_1 > 1$ (i.e., when there are coded categorical predictor variables), the robust distances in equation 8 are appropriate as a leverage diagnostic but not (in the author's opinion) as a distance measure in a multivariate setting. Therefore we recommend that software report the leverage diagnostic on the scale of the hat values.

4 Example

We turn to the epilepsy data published in Thall and Vail [1990] for an example.

```
> require(robustbase)
> data(epilepsy)
```

First make the design matrix.

```
> X <- model.matrix(~ Age10 + Base4 * Trt, data = epilepsy)
> n <- nrow(X)
> head(X)

  (Intercept) Age10 Base4 Trtprogabide Base4:Trtprogabide
1           1   3.1  2.75            0                  0
2           1   3.0  2.75            0                  0
3           1   2.5  1.50            0                  0
4           1   3.6  2.00            0                  0
5           1   2.2 16.50            0                  0
6           1   2.9  6.75            0                  0
```

In this case we have

```
> X1 <- X[, c(1, 4)]
> head(X1)

  (Intercept) Trtprogabide
1           1            0
```

```
2           1           0
3           1           0
4           1           0
5           1           0
6           1           0
```

```
> X2 <- X[, 2:3]
> head(X2)
```

```
  Age10 Base4
1   3.1  2.75
2   3.0  2.75
3   2.5  1.50
4   3.6  2.00
5   2.2 16.50
6   2.9  6.75
```

```
> X3 <- X[, 5, drop = FALSE]
> head(X3)
```

```
  Base4:Trtprogabide
1                  0
2                  0
3                  0
4                  0
5                  0
6                  0
```

```
> mcd <- covMcd(X2)
> w <- mcd$raw.weights
> mcd$cov
```

```
           Age10      Base4
Age10  0.7463740 -0.3267283
Base4 -0.3267283 10.0194113
```

The implementation of the Fast MCD in the robustbase package rescales the final covariance matrix estimate by a consistency correction factor `mcd$cnp[1]` and a small sample correction factor `mcd$cnp[1]` so that $c =$ `prod(mcd$cnp)`.

```
> cov.wt(X2, wt = w)$cov * prod(mcd$cnp)
```

```
           Age10      Base4
Age10  0.7463740 -0.3267283
Base4 -0.3267283 10.0194113
```

```
> TX2 <- apply(X2, 2, weighted.mean, w = w)
```

Compute $\tilde{X}_2$ by applying equation 5 to X_2.

```
> X2.tilde <- sweep(X2, 2, TX2)
> X2.tilde <- sqrt(prod(mcd$cnp)*(n - 1)/(sum(w) - 1) * w) * X2.tilde
> X2.tilde <- sweep(X2.tilde, 2, TX2, FUN = "+")
```

Verify that $C(\tilde{X}_2) = C^{(rob)}(X_2)$.

```
> var(X2.tilde)
```

```
           Age10      Base4
Age10  0.7463740 -0.3267283
Base4 -0.3267283 10.0194113
```

We can obtain the modified data (not in general but for this example) by replacing X_2 in the original data and recomputing the design matrix.

```
> epilepsy[dimnames(X2)[[2]]] <- X2
> X.tilde <- model.matrix(~ Age10 + Base4 * Trt, data = epilepsy)
> head(X.tilde)
```

```
  (Intercept) Age10 Base4 Trtprogabide Base4:Trtprogabide
1           1   3.1  2.75            0                  0
2           1   3.0  2.75            0                  0
3           1   2.5  1.50            0                  0
4           1   3.6  2.00            0                  0
5           1   2.2 16.50            0                  0
6           1   2.9  6.75            0                  0
```

The final robust leverage measure is then given be the diagonal element of the matrix

$$X(\tilde{X}^\top \tilde{X})^{-1} X^\top.$$

```
> diag(X %*% solve(t(X.tilde) %*% X.tilde) %*% t(X))
```

```
         1          2          3          4          5          6          7
0.05918398 0.05761964 0.07597885 0.08831037 0.12814167 0.03649363 0.05707197
         8          9         10         11         12         13         14
0.13821982 0.06977150 0.05953140 0.08231479 0.04790064 0.06109578 0.06518841
        15         16         17         18         19         20         21
0.21304208 0.06047114 0.04498471 0.38633944 0.04914452 0.07172279 0.05490496
        22         23         24         25         26         27         28
0.09056742 0.05061124 0.04363259 0.06789648 0.12056569 0.10505741 0.07403980
```

```
        29         30         31         32         33         34         35
0.13316337 0.04489245 0.07575642 0.05223374 0.09433237 0.04382864 0.03457940
        36         37         38         39         40         41         42
0.06124138 0.05326251 0.09628077 0.04761239 0.05961493 0.05079567 0.10109938
        43         44         45         46         47         48         49
0.06090713 0.05230413 0.06278511 0.06904524 0.03396855 0.05985715 0.64794379
        50         51         52         53         54         55         56
0.04181870 0.03780989 0.05743717 0.06796775 0.11009718 0.04673072 0.03927901
        57         58         59
0.05935622 0.06818611 0.07601004
```

References

Ricardo A. Maronna and Victor J. Yohai. Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89(1–2):197–214, 2000. ISSN 0378-3758. doi: 10.1016/S0378-3758(99)00208-6. URL `http://www.sciencedirect.com/science/article/pii/S0378375899002086`.

Peter J. Rousseeuw and Katrien van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41 (3):212–223, 1999. doi: 10.1080/00401706.1999.10485670. URL `http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1999.10485670`.

Peter J. Rousseeuw and Bert C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):pp. 633–639, 1990. ISSN 01621459. URL `http://www.jstor.org/stable/2289995`.

P.F. Thall and S.C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, pages 657–671, 1990.